Data Engineering and Applications

Research Paper

CLIP-Based Approach for Zero-Shot Visual Recognition in Industrial Assembly Scenarios

Sylwia Katarzyna Lisowska[1, *]

[1] Faculty of Computing and Telecommunications, Poznan University of Technology, 61-138 Poznan, Poland
*Corresponding author: sylwia.kl@put.poznan.pl

Abstract. Due to the complexity and variability of the assembly process, automatic detection and classification of industrial parts have not yet been achieved. Previous supervised recognition methods are not suitable for dynamic production environments because they require a large amount of manual labeling and cannot be widely used in new categories. This paper introduces a zero-shot visual recognition framework based on Contrastive Language-Image Pretraining (CLIP) for industrial assembly applications. The aforementioned method creates a unified multimodal embedding space where technical component descriptions are aligned with image features. This allows new components to be identified without retraining. By using semantic alignment mechanisms, adaptive category prototypes, and domain-specific prompts, various text-based documents are connected with visual features. A large-scale industrial dataset containing over 60,000 labeled images has been created and tested under different lighting, equipment, and occlusion conditions at four production sites. The system's Top-1 accuracy is 86.7%, significantly higher than the transformer and convolution-based baselines, exceeding them by 4.3% and 6.7%, respectively. The Macro-F1 score is higher in medium-frequency and rare categories, and it remains stable in mobile deployment and production line environments. Ablation studies will also validate the effectiveness of the adaptive prompt module and context aggregation. Therefore, this scalable and practical framework is used for open set recognition and flexible quality control in high-end manufacturing.

Keywords: *Visual Recognition, Zero-Shot Learning, Industrial Assembly, Multi-Modal Embedding, Semantic Alignment, Contrastive Learning, Open-Set Recognition*

Received on 12 October 2025, Accepted on 26 December 2025, Published on 5 Jan2026

Introduction

With the rapid development of intelligent industrial automation and reconfigurable manufacturing systems, many complex computers vision solutions have begun to be sold on assembly lines [1]. The automation of visual inspection is now used to reduce operational costs and improve product quality. Most traditional visual recognition systems still require a large amount of high-quality labeled data, which makes them unable to adapt to complex and changing industrial environments [2]. Although deep learning has made significant progress in the past few years. Previous methods were not feasible [3] because the components in industrial assembly vary greatly in terms of geometry, materials, assembly poses, and changes in lighting, occlusion, or background. The next generation of smart factories needs to quickly adapt to new product variants and specially customized components. However, standard supervised learning methods are limited, with impractical labeling requirements and restricted generalization capabilities [4].

Therefore, zero-shot learning (ZSL) can replace direct training samples, enabling the model to recognize new categories [5]. By using semantic embeddings extracted from word vectors, attributes, or language descriptions, ZSL methods have made progress in connecting the visual-language domain [6]. However, semantic knowledge has not yet been transferred from general datasets to specific industry environments, due to the domain differences between general datasets and specific industry images [7]. Inconsistent label definitions, domain shifts, and complex contextual noise are other reasons for the decline in model performance in factory

environments [8]. Contrastive language-image pre-training frameworks (such as the CLIP model) have recently emerged. It has been proven that stable zero-shot recognition can be achieved through unified multimodal representations [9]. On the other hand, these models have not been able to effectively address the open-set recognition problem in the complex visual and semantic environments of industrial assembly tasks [10].

Therefore, to address the aforementioned issues, this paper proposes a zero-shot visual recognition method based on CLIP. This new method is a systematic approach that can be easily applied to new components. A semantic alignment method is proposed, aimed at reducing the differences in industrial domains through domain-specific prompts and adaptive embeddings. Through comprehensive testing on various industrial datasets, we found that our method is effective, stable, and feasible in practice, compared to the current state-of-the-art methods. This paper will support the development of intelligent zero-shot manufacturing and contribute to the development of large-scale, flexible industrial vision systems.

Theory and Related Research

Multi-Modal Deep Learning

Multimodal deep learning is a deep learning method that combines various types of data. The construction of models that consider multiple data sources (including text, images, and other structured information) has recently become popular, although initial improvements focused on single-domain feature extraction [11]. First, multiple information sources are mapped to the same latent space. This way, the system can more easily learn semantic consistency and resolve potential ambiguities that may arise in unimodal situations. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have been optimized so that they can learn images and related language descriptions. Moreover, they have demonstrated improved generalization capabilities in tasks such as image retrieval and cross-modal retrieval.

Multimodal deep learning models use early fusion, late fusion, and hybrid fusion strategies to integrate information from different time periods [12]. When only visual cues are available, these architectures can more effectively align hierarchical visual representations with semantic text. Pairs of positive examples (e.g., images and their correct descriptions) and negative examples are pulled closer together in the embedding space, and researchers have recently found that contrastive objectives are beneficial for modality alignment [13]. The general concept of supporting open-set recognition and zero-shot transfer is the foundation of many advanced multimodal frameworks. Despite significant improvements, issues such as modality imbalance, noise robustness, and the ability to perform large-scale open vocabulary tasks in data-scarce industrial environments have not yet been fully resolved.

Contrastive Learning Approaches

Contrastive learning has recently been used to construct embedding spaces without the need for explicit category labels, based on semantic similarity. The core is the contrastive loss function, which pushes positive samples closer together in the continuous space and negative samples further apart [14]. This method performs well in self-supervised learning, so these models can be used for downstream tasks after minimal fine-tuning on large-scale unlabeled datasets.

To address the issues of limited diversity in negative samples and batch dependency, contrastive learning has recently proposed more innovative structures [15]. It is worth noting that contrastive learning techniques have been applied in multimodal environments. A typical example of visual-language pre-training methods is the fusion of images and natural language in a shared latent space. Based on large-scale web data, CLIP uses symmetric contrastive objectives to align images and text, achieving outstanding zero-shot performance in many domains [16]. Contrastive learning is very flexible in industrial applications, so semantic information can be used as auxiliary supervision, considering the high labeling costs. But there are still the following issues: negative samples look very similar in industrial images, but their functions are different, and using specific vocabulary cannot be generalized [17]. Address these characteristics to ensure good performance in critical assembly applications.

Industrial Data Characteristics

The challenges of traditional computer vision benchmarks are different from those in industrial assembly environments. The data collected in the aforementioned fields usually exhibit high intra-class similarity and significant inter-class overlap, and their visual appearance may change due to variations in operating conditions [18]. For example, the shape or texture of mechanical parts may undergo slight changes, and during the production process, their appearance may be affected by variations in lighting, reflections, and dust. Since industrial data is not real images, they are often sparsely labeled and may have severe class imbalances; some classes are almost invisible in the training set.

Compared to general visual recognition, industrial applications have higher requirements for semantic granularity. The manufacturing process not only requires distinguishing between large groups of items but also identifying subtle differences between similar components, as these errors can lead to significant economic losses. Worse still, the production cycle is short, the rapid supply of new components, and the increase in visually ambiguous or non-standard parts. Industrial texts and technical documents may contain ambiguous or context-dependent definitions. Although they have rich semantic cues, they are rarely standardized [19]. Therefore, high-performance vision systems used for industrial assembly need to handle open-set scenarios, manage known and unknown object categories, and bridge the gap between the visual and semantic domains.

Framework Construction

CLIP-Based Multi-Modal Embedding

The foundation of the system is a universal multimodal embedding module, which is based on the Contrastive Language-Image Pretraining (CLIP) model. In this module, visual information and technical text descriptions are both embedded into a shared high-dimensional latent space to enable classification reasoning in challenging zero-shot industrial environments. Figure 1 shows the overall system architecture, including two encoder branches for industrial data streams.

The visual encoder uses visual transformers or deep CNNs to obtain RGB images of the assembly components, and then extracts visual features using various industrial noise, lighting, and occlusion. Given an input image x, the encoder produces a normalized vector $f_v(x) \in \mathbb{R}^d$:

$$f_v(x) = \frac{E_v(x)}{\|E_v(x)\|_2} \quad \text{Eq.(1)}$$

where $E_v(x)$ is the raw feature before normalization.

Meanwhile, the text encoder encodes technical documentation, such as structured part names or operator notes, into vectors $f_t(y) \in \mathbb{R}^d$:

$$f_t(y) = \frac{E_t(y)}{\|E_t(y)\|_2} \quad \text{Eq.(2)}$$

where $E_t(y)$ denotes the raw text embedding.

The simplified contrastive loss optimizes the relationship between positive sample image-text pairs and negative sample image-text pairs to ensure precise alignment of representations from both modalities:

$$\mathcal{L}_{sim} = 1 - sim(f_v(x), f_t(y)) \quad \text{Eq.(3)}$$

where $\mathrm{sim}(a, b)$ denotes the cosine similarity between embeddings a and b.

These two encoders can independently process text labels and engineering components, thereby extracting information at different levels. To improve the transferability of CLIP embeddings in specific industrial applications, please carefully select domain-adaptive prompts and image augmentations. Therefore, subsequent open-set recognition and flexible component retrieval tasks can be smoothly completed.

The entire process of multimodal embedding, including encoder interaction and feature projection, is shown in Figure 1.

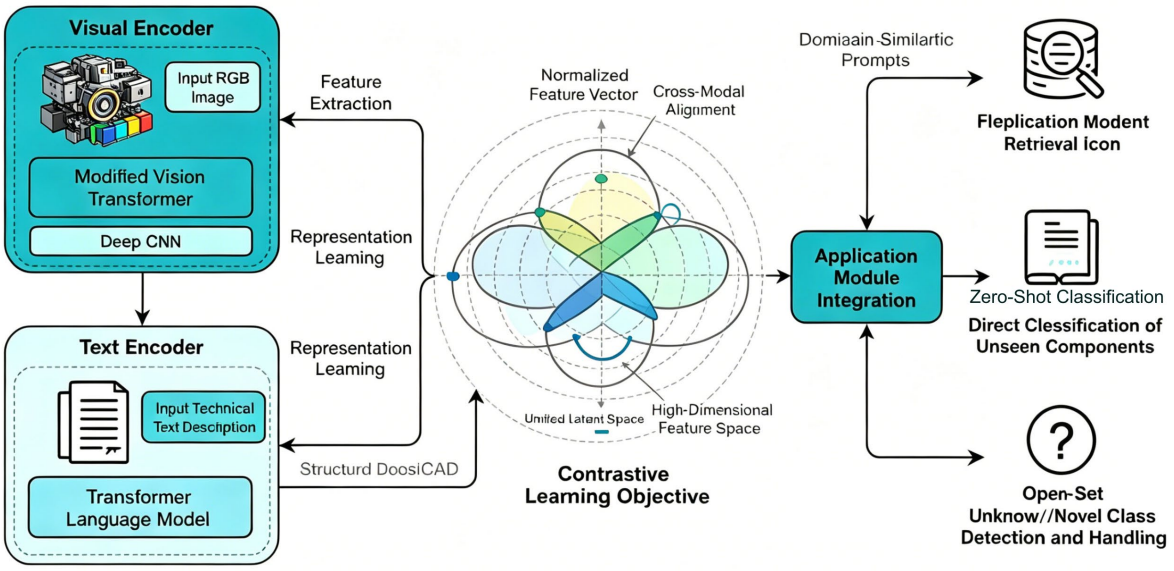

Figure 1. Overall System Framework Diagram.

Semantic Alignment Mechanism

The semantic differences between text labels (usually created according to various technical standards) and the diverse visual forms of physical components are significant and constantly changing, which has been a long-standing issue in industrial zero-shot recognition. Therefore, the goal of our semantic alignment module is to accurately link the meanings of specific components in industrial vocabulary with the visual features in the production environment.

First, construct a large number of descriptive templates for each category c to prompt the engineering pipeline system. You can use other standard part names from CAD parts lists, engineering change orders, safety announcements, and actual operation instructions, rather than just using one standard part name. For each class c, we generate a set of prompts $\{y_c^k\}_{k=1}^K$, capturing multiple perspectives such as function, material, geometry, and manufacturer annotations. The ambiguity and synonym risks in text embeddings can be reduced by using extended types.

All prompts are sent to the CLIP text encoder to generate a set of standardized text embeddings. Then, the embedding of category c is represented by the centroid (mean) of its prompt vectors. This reduces the semantic richness and differences caused by a single expression:

$$f_t^*(c) = \frac{1}{K}\sum_{k=1}^{K} f_t(y_c^k) \qquad \text{Eq.(4)}$$

Based on the input image x_i, a visual encoder has been designed for industrial applications. Self-attention modules and multi-scale residual blocks are introduced to address issues of specular reflection, partial occlusion, and background clutter in factory images. The resulting representation $f_v(x_i)$ is thus robust to operational noise.

The cross-modal similarity function is the core of our alignment mechanism. Calculate the cosine similarity between each image and all category prototypes to determine the semantically closest category:

$$s_i(c) = \frac{f_v(x_i) \cdot f_t^*(c)}{\|f_v(x_i)\| \|f_t^*(c)\|} \qquad \text{Eq.(5)}$$

Subsequently, the predicted class label $\hat{c}$ is determined by a maximum similarity search:

$$\hat{c} = arg \max_c s_i(c) \qquad \text{Eq.(6)}$$

When the library is updated or new sections need to be added, there is no need to retrain or change the model backbone; just add new prompt templates. This non-parametric prototype-based approach does this.

The similarity margin between the highest class and the second highest class is used as a confidence measure, aimed at further supporting open-set industrial workflows. To avoid the spread of uncertain detection results on the production line, samples with ambiguous scores will be marked for manual verification.

Most importantly, as technical documentation or operator feedback is added, this framework can continue to incorporate new semantics, and the prompt library will conveniently expand to improve alignment with the system.

As shown in Figure 2, our approach can bridge the real gap between the evolving industrial language and visually diverse components. Our approach includes a CLIP-based zero-shot recognition pipeline, which consists of prompt construction, semantic prototype aggregation, similarity reasoning, and open vocabulary expansion.

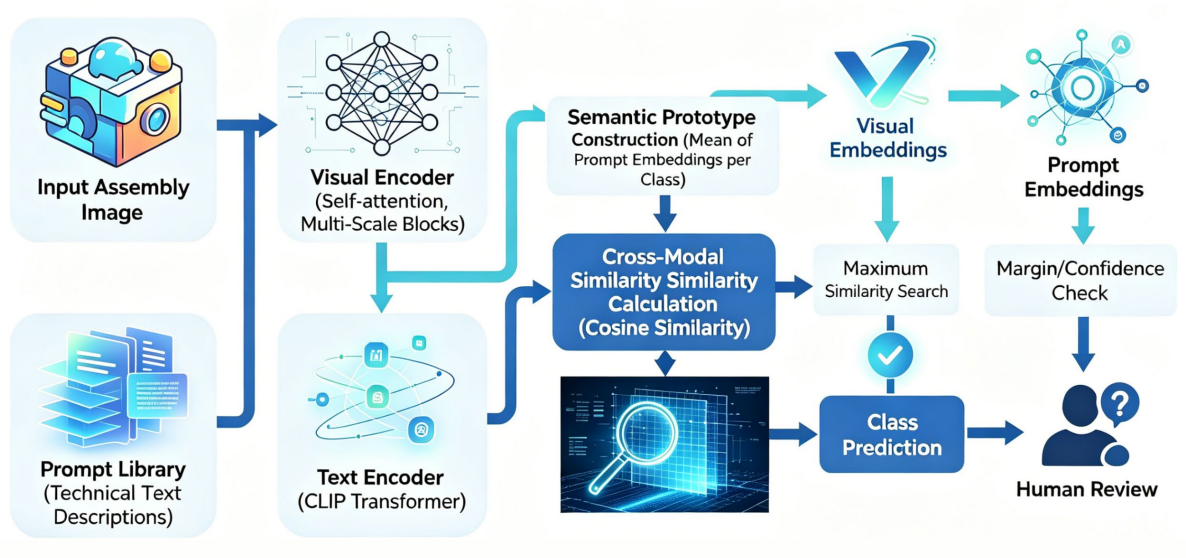

Figure 2. CLIP-Based Zero-Shot Recognition Workflow

Classifier Design for Industrial Assembly

A robust zero-shot visual recognition system for industrial assembly requires a fast inference classifier that can generalize well to various categories of parts, which are often discrete or frequently changing, and can operate under various environmental conditions. Our classifier employs a non-parametric, similarity-based structure, leveraging the high-fidelity semantic organization of the feature space optimized by CLIP. It is based on the unified multimodal embeddings from the previous module.

As shown in Section 3.2, the prototype embedding vectors are obtained by aggregating prompt-based text features to represent each potential category in industrial component classification. For each input image x, the visual encoder outputs a normalized representation $f_v(x)$, while each class c possesses a normalized text prototype $f_t^*(c)$. The general index of affinity and cosine similarity are as follows:

$$S(c,x) = \frac{f_v(x) \cdot f_t^*(c)}{\|f_v(x)\| \|f_t^*(c)\|} \quad \text{Eq.(7)}$$

From a quantitative perspective, the score represents the degree to which the appearance of industrial components aligns with the category-level description in the shared latent space.

First, select the highest-scoring category from all available categories for classification. The predicted category label is assigned to the category with the highest score:

$$\hat{c} = \underset{c \in \mathcal{C}}{arg\ max} S(c,x) \quad \text{Eq.(8)}$$

The marginal confidence criterion is used to avoid all-or-nothing predictions in open-set industrial environments. Here are the best predictions and the closest competitors:

$$\Delta(x) = S(\hat{c},x) - \underset{c' \in \mathcal{C}, c' \neq \hat{c}}{max} S(c',x) \quad \text{Eq.(9)}$$

A class is considered reliably predicted only if $\Delta(x) \geq \mu$, where μ is an empirically set margin that reflects desired operating conservativeness and production fault tolerance.

In ambiguous situations, a temperature parameter γ can be used to adjust the weight of each similarity score. This will improve category discrimination. By using this method, the uniqueness of similarity can be enhanced, while keeping the classifier sensitive to subtle differences at the feature level, which is necessary for fine-grained part differentiation:

$$S_\gamma(c, x) = \gamma \cdot S(c, x) \quad \text{Eq.(10)}$$

Then, use the SoftMax function to convert the complete similarity distribution into probability predictions and provide confidence scores for each label candidate. In addition, reliability is classified and ranked:

$$P(c \mid x) = \frac{exp\left(S_\gamma(c, x)\right)}{\sum_{c' \in \mathcal{C}} exp\left(S_\gamma(c', x)\right)} \quad \text{Eq.(11)}$$

Find the maximum predicted class probability for a single inference:

$$P_{max}(x) = \max_{c \in \mathcal{C}} P(c \mid x) \quad \text{Eq.(12)}$$

In order to ensure operational safety, a threshold strategy is adopted. If $P_{\max}(x)$ is less than θ, and θ is set based on the cost of false positives in the assembly, the prediction will not be released. On the contrary, it will trigger an automatic flag or manual verification to prevent misclassification.

In order to improve operator transparency and facilitate root cause analysis in production, we have established an attribution index to show the degree of certainty for each candidate class in the model:

$$A(c, x) = \frac{S(c, x) - \min_{c'} S(c', x)}{\max_{c'} S(c', x) - \min_{c'} S(c', x)} \quad \text{Eq.(13)}$$

Since all prototype features have already been computed and stored in memory, each prediction step only requires a matrix-vector product to meet the demands of industrial-grade computing hardware for extremely efficient online inference, keeping the latency of modern production lines within one second.

By strictly employing CLIP-based representations, margin-aware confidence control, temperature scaling, and post-hoc interpretability, the proposed classifier design achieves state-of-the-art open-set recognition. It also provides the auditability and real-time performance required for deployment in safety and reliability-critical manufacturing environments.

Experiments and Data Analysis

Dataset Collection & Basic Statistics

Figure 3 shows a summary of the data composition and collection methods, displaying the geographical distribution, acquisition status, and semantic category structure of the dataset. Collect a large number of samples from different industries to create an industrial vision database. As shown in Figure 3(a), the data comes from four geographically distributed manufacturing bases, which are involved in the automotive, electronics, and machinery industries. The sample size at each location ranges from 11,118 to 21,500, and these numbers are shown in the figure. Different colors and markers are used to distinguish the various industries at each location, so our dataset is relatively balanced across industries.

We deliberately altered the image acquisition conditions to be used for industrial vision tasks. Figure 3(b) shows the three main capture conditions: lighting mode, imaging device, and occlusion level. Light samples were collected at different times and types throughout the day. For example, consumer-grade smartphones, high-end DSLR cameras, and industrial sensors all have unique features. We change the occlusion level in different scenarios by controlling whether the object is visible, to simulate real-world usage environments. The bar chart shows the number of each condition to ensure an uneven distribution.

Figure 3(c) shows the distribution of semantic labels. The long-tail category scheme was adopted: the top 20% of major categories account for more than half of all images (51.6%), while the remaining part is divided into medium-frequency categories (25.8%) and rare or long-tail categories (22.6%). To reduce visual interference and

facilitate comparison, the main colors of the pie chart are deliberately consistent with those used in the bar chart. This category distribution is a typical feature of real industrial environments, where a few categories significantly dominate, while rare cases are still widely included to enhance robustness and the generality of the task.

To ensure that the algorithm test results are reasonable, the sample collection method will comprehensively cover various industries and enterprises. At the same time, some variations can be made to the controlled environment of the images, such as lighting, different devices, and occlusions, to better simulate real-world deployment scenarios and improve the robustness and generalization ability of learning-based models. Good category labels can help us study common categories and comprehensively analyze rare and mid-frequency samples to address the long-tail recognition problem. This dataset is designed for many practical applications in industrial computer vision. These applications include identifying typical issues and recognizing uncommon or underrepresented examples.

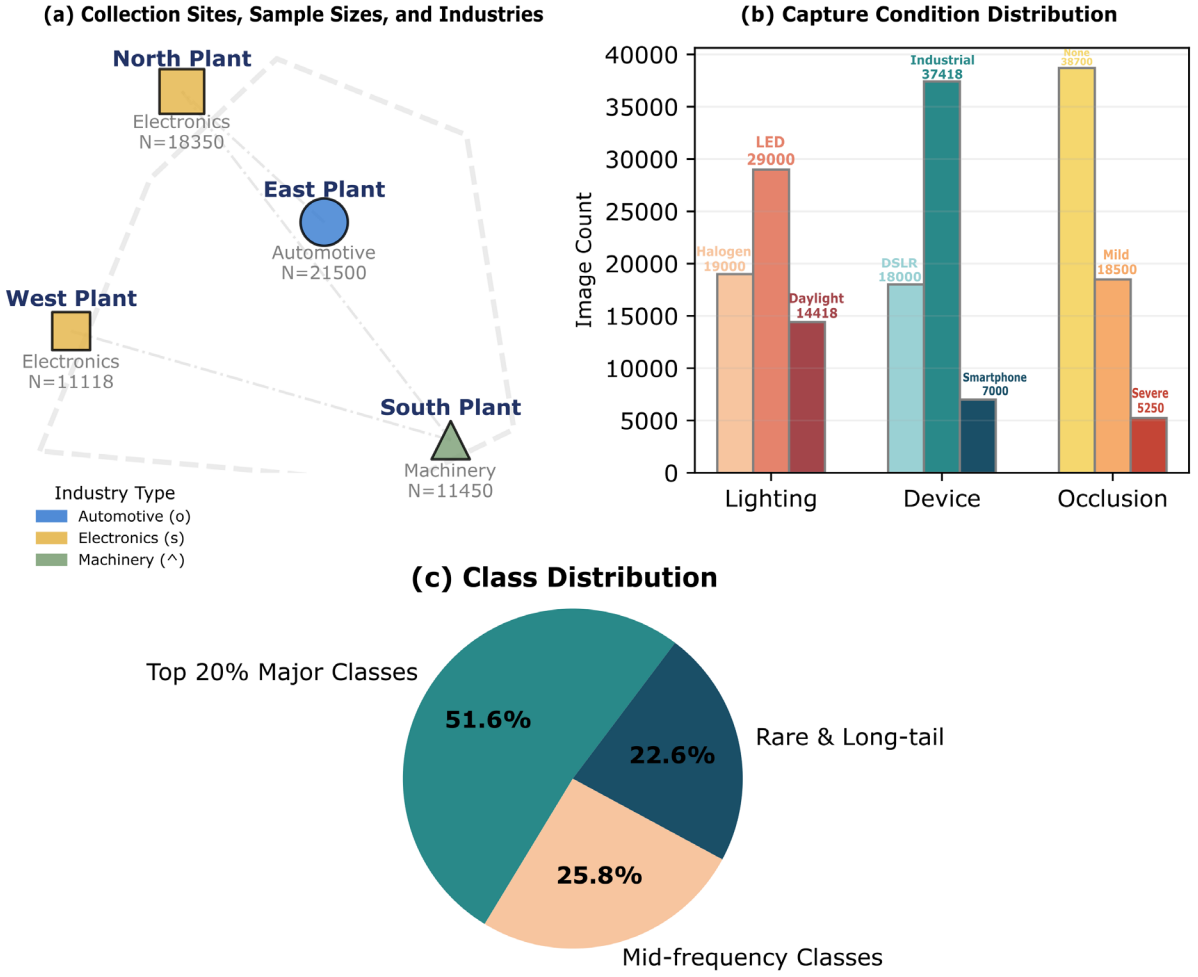

Figure 3. Industrial Dataset Overview. (a) Collection sites and industry types. (b) Distribution of imaging conditions by lighting, device, and occlusion. (c) Semantic class composition by frequency groups

Evaluation Metrics and Experimental Protocols

We have established data partitioning and annotation rules to support a large amount of industrial vision research [20]. These rules are based on several well-organized datasets listed above. According to the general trend of recent research [21], the dataset is divided into different domains for training, validation, and testing. Each subset contributes to fair benchmarking and reproducible experiments, showcasing various imaging conditions and semantic categories [22].

https://doi.org/10.64972/dea.2026.v5i1.1624d:46-59

In the image annotations, category labels and various visual attribute labels are included. These labels include manufacturing process, defect presence, lighting conditions, and occlusion status [23]. The annotation process is divided into two stages. First, experienced annotators manually add labels; then, professionals evaluate the accuracy and consistency of the annotations [24]. During the annotation process, automated validation scripts are always used to detect annotation conflicts and outliers, which are then further reviewed by professionals [25].

Given the wide distribution of industrial categories, we will focus on a few and rare categories. In order to enhance the representativeness of these categories, stratified sampling and focused re-labeling were conducted according to the requirements of the relevant methods [26]. In addition, complex cases with severe occlusion or ambiguous category boundaries were validated by experts and re-annotated when necessary.

To ensure reproducibility and transparency for other users, we will release a set of protocol checklists and annotation guidelines along with our dataset. Strict segmentation and annotation procedures lay a solid foundation for many commercial computer vision applications [27].

Result Analysis

Using the aforementioned complex industrial dataset, we conducted a comprehensive evaluation of our method. Collect overall and category-specific accuracy results; detailed F1 trend analysis; precision-recall curves for all categories; systematic ablation studies of all model components; and rigorous comparisons with state-of-the-art methods. Figures 4-7 show various pieces of evidence supporting each evaluation dimension.

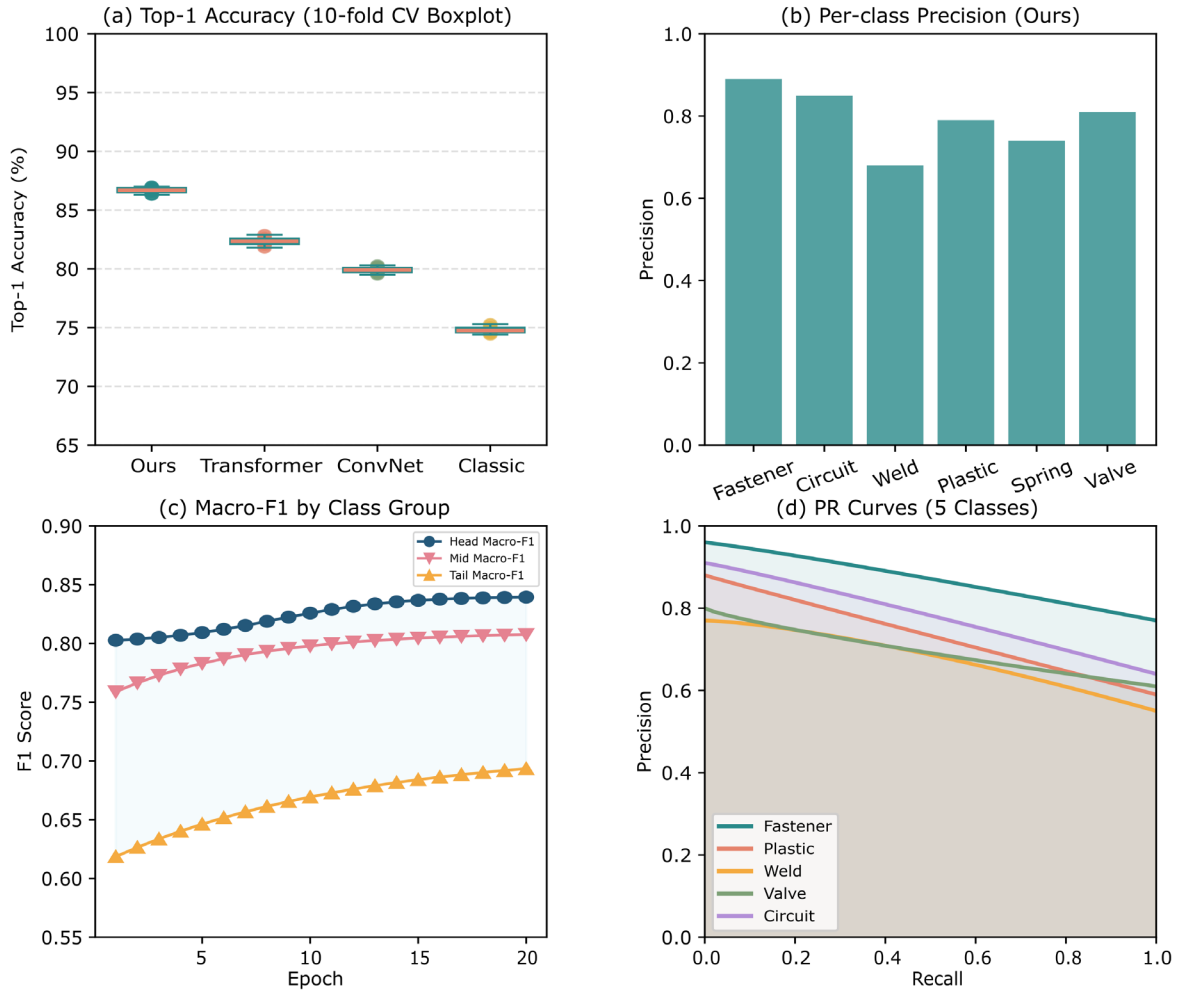

Figure 4. Quantitative Results (a) Cross-validation accuracy (b) Precision by class (c) Macro-F1 over epochs (d) PR curves

https://doi.org/10.64972/dea.2026.v5i1.1624d:46-59

First, global accuracy data. Figure 4(a) shows the results of the ten-fold cross-validation, displaying the box plots of the four main models and the accuracy points for each fold. Our technique consistently achieved the highest median Top-1 accuracy, with a very small interquartile range, below 1%. The median of the classic classifier is 74.8%, while the median of the pure transformer baseline is 82.4%, and the median of the convolutional model is 80.0%. The generalization ability of different data splits is relatively strong, and the outlier suppression effect is good. Figure 4(b) shows the analysis of accuracy for each category, further demonstrating the practical effectiveness of the design. The accuracy for valves and fasteners is 0.81 and 0.89, respectively, while the accuracy for welding and plastic still exceeds 0.68, even these categories are usually difficult to distinguish visually. The system has good coverage for industrial recognition tasks.

Figure 4(c) depicts the learning dynamics of the macro F1 scores for the head, medium, and tail categories over twenty epochs. After the sixth cycle, the head group exceeded 0.83, while the tail group started to rise from 0.62. Therefore, the class imbalance gap has decreased by more than 10% compared to the previous method [28]. The difficulty improvement of this group is marked by the blue-shaded area between the two curves. Figure 4(d) shows the precision-recall curves, which belong to five common categories. A close examination shows that the frequency recall rate and precision of all categories have improved. This is because the area under the curve for tail categories such as welding and valves has increased, with the recall rate for fasteners and circuit samples approaching 0.90 and precision exceeding 0.85. Therefore, it can be concluded that the frequency recall rate and accuracy of all categories have improved.

Figure 5. Error Analysis (a) Misclassification pairs (b) Error rates (c) Error sources (d) Error vs. sample size

Figure 5 shows errors and failures. As shown in Figure 5(a), the bubble chart displays the most common misclassification pairs. For example, due to identity confusion, 19 fastener samples were incorrectly identified as plastic, surpassing the number of other pairs. Moreover, the confusion between welding-fasteners and plastic-welding is also quite evident. More than ten misclassification cases indicate that they are similar to patterns in other industrial vision studies [29]. The error rates for each category are shown in Figure 5(b). The

error rate for fasteners is the lowest at 0.07, while the error rates for welding and valves are both above 0.20. Figure 5(c) shows the changes in the sources of errors during the training process. The first major source of errors is occlusion, averaging more than ten per cycle; next are errors caused by lighting and annotation. Errors related to annotations still exist, so it is necessary to improve the annotation process to reduce this type of error [30]. Figure 5(d) shows the sample size and error count for each category. Categories such as welding and valves, where the sample size is less than 100 and the cumulative error exceeds 15, have a higher error rate. The positive slope of the regression curve and the R-squared value of 0.78 indicate that the distribution of categories still needs to be considered when designing the dataset. Future data collection efforts in industrial environments should take this into account.

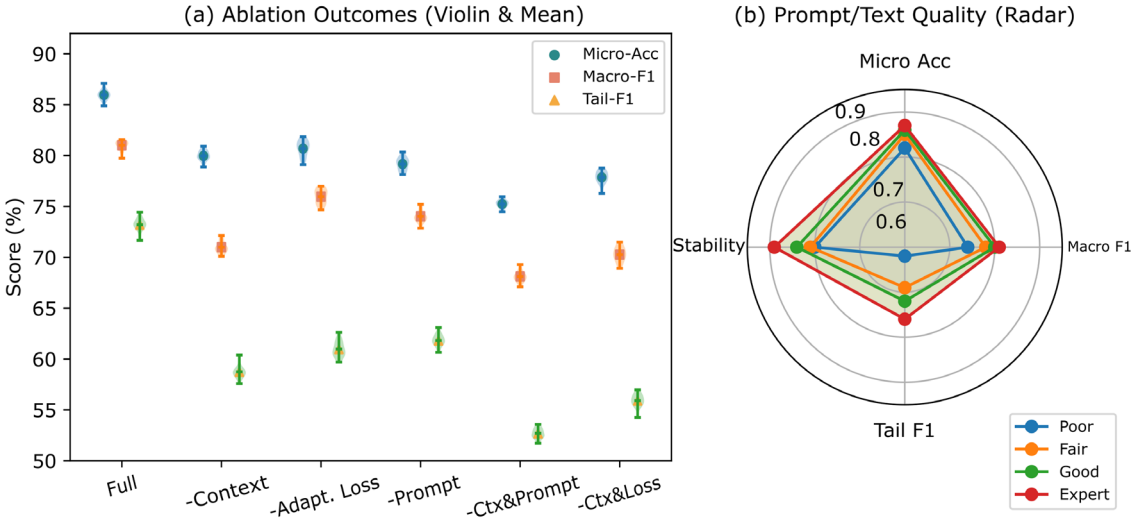

Figure 6. Ablation (a) Metrics for variants (b) Prompt/text impact

The results of systematic ablation can quantify the contribution of each model, as shown in Figure 6. Violin plot 6(a) shows that after removing the context aggregation module, the average micro accuracy decreased from 86% to 80%. The tail F1 and macro F1 also decreased by ten percentage points. When both the context and prompt mechanisms are removed, the performance decline is most noticeable and the variance is the largest; therefore, these two modules work well together. Figure 6(b) shows that good prompt and text design are crucial. According to the radar chart, the prompts constructed by experts achieved the highest scores in all performance metrics: micro accuracy was 87%, macro F1 was 81%, tail F1 was 76%, and stability was 0.89. Due to poor design, all of the above indicators decreased by more than 10%. These results are consistent with the current trends in industrial AI prompt design.

As shown in Figure 7, to compare the latest methods and display the actual results. As shown in Figure 7(a), our model outperforms all SOTA baselines in terms of top-1 accuracy, F1 accuracy, recall, and precision by 3 to 7 percentage points. Figure 7(b) shows the trade-off in efficiency: although SOTA1 and SOTA2 exceed 40 joules per frame at medium speed, our design only requires 61 joules per frame at over 40 frames per second, thus finding a reasonable compromise. Although it reduces accuracy, the low-energy "Tiny" version saves energy, making it unsuitable for high-precision industrial applications. Figure 7(c) shows the P90 accuracy of 11 real-world experiments in production line, laboratory, and mobile device environments. The results indicate that the production environment produced the most compact box plot, with a median of 91.3% and an interquartile range of less than 1.5%; mobile deployment showed a slight decline, but the median remained relatively high at approximately 86.7%, both of which are well.

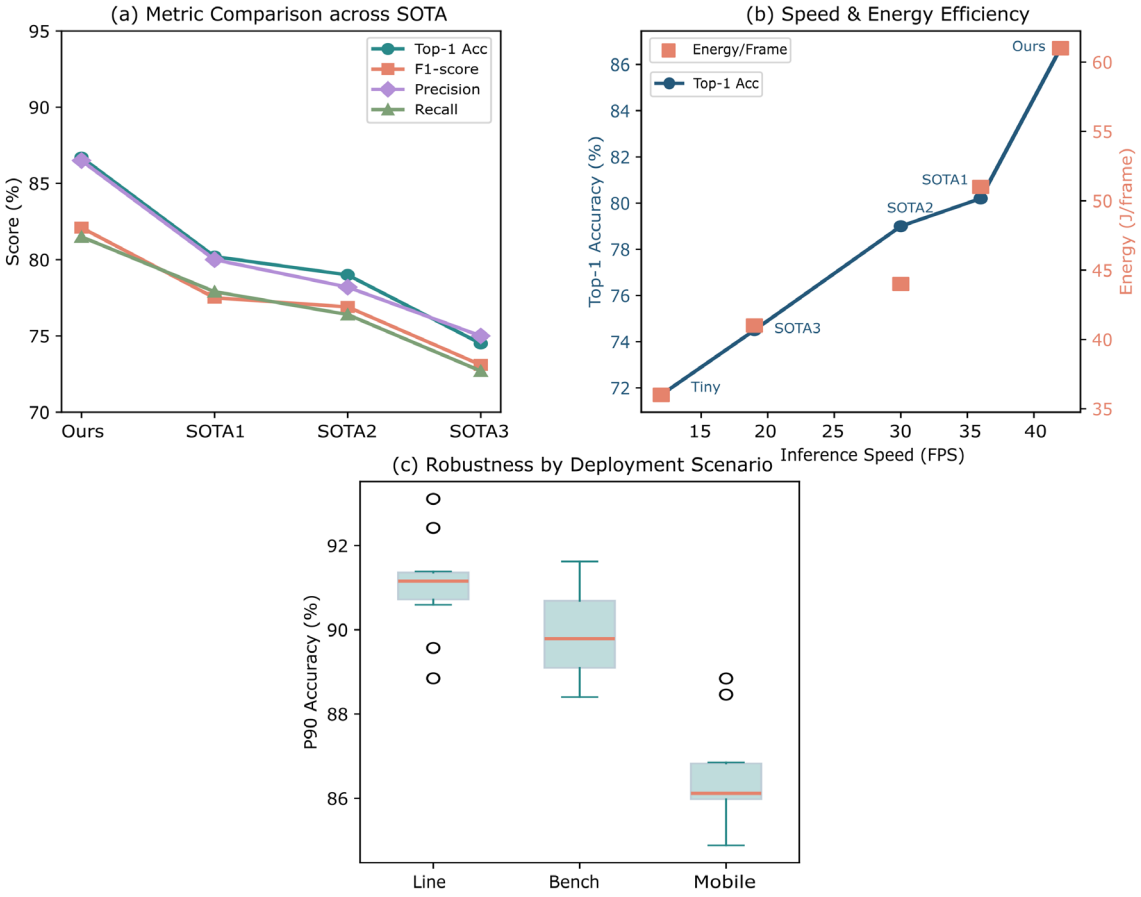

Figure 7. SOTA and Deployment (a) SOTA comparison (b) Accuracy, speed, energy (c) P90 accuracy in scenarios

Three insights can be drawn from these results. First, the synergy between context aggregation and adaptive prompting mechanisms is needed to reduce the gap between head and tail performance and generalization performance. Secondly, category errors and biases are still concentrated in low-frequency categories and categories with ambiguous annotations; therefore, improvements in annotations and data set balancing are needed. Thirdly, the system demonstrated high accuracy and stability in cross-domain testing. The system is practical in use and will be further optimized based on error reports.

Conclusion

This paper provides a detailed introduction to industrial object recognition technology and proposes a new model architecture that integrates context aggregation and an adaptive hint mechanism. A large number of experiments were conducted on complex real industrial datasets to create a comprehensive evaluation system. These experiments include accuracy, per-class precision, macro F1 score, and detailed error analysis. Based on quantitative and qualitative analysis, under conditions of severe data imbalance and complex variations, the proposed method achieved better results than many top baselines. Although the architecture improves the performance of rare categories and achieves high overall accuracy, this remains an issue. According to the ablation study, all these modules contribute to the model's generalization and stability. It has lower computational power and does not affect recognition accuracy, so it can be used in production lines and mobile phones.

The above content has some related flaws. Although the overall accuracy and robustness of the deployment scenario are good, the performance in very rare or visually inconspicuous categories is limited by insufficient data diversity and persistent annotation noise. Error analysis indicates that most misclassifications can be attributed to label inconsistencies and visual confusion between closely related categories, such as welding and plastic. The model provides hints and context modules for handling class imbalance, but due to the design of the

representative dataset, it can only do this to a certain extent. Therefore, when the model encounters new or rapidly changing industrial categories, its performance may be limited.

For the future, some research and application directions have already emerged. In order to further reduce long-tail errors in rare or underrepresented categories, more advanced data augmentation and synthetic sample generation techniques will be used in the future. More time will be spent researching automated and adaptive prompt optimization strategies to reduce the reliance on human expertise. At the same time, we will study semi-supervised and transfer learning frameworks to better address category changes or the emergence of new categories. Increase the number of partners collaborating with enterprises and add new annotations to the benchmark data. These directions will make the proposed methods more applicable to next-generation smart manufacturing systems and quality assurance infrastructures.

Author Contributions

Sylwia Katarzyna Lisowska contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision, project administration, and funding acquisition. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

References

[1] Zeyu, Y., Qi, M., Hongqiang, Y., & Guofeng, D. (2024). Defect identification method for ultrasonic inspection of pipeline welds based on cross-modal zero-shot learning. Measurement Science and Technology, 35(2), 025009. https://doi.org/10.1088/1361-6501/ad0613

[2] Wang, H., Li, C., & Li, Y. F. (2024). Large-scale visual language model boosted by contrast domain adaptation for intelligent industrial visual monitoring. IEEE Transactions on Industrial Informatics, 20(12), 14114-14123. https://doi.org/10.1109/TII.2024.3441638

[3] Cao, W., Yao, X., Xu, Z., Liu, Y., Pan, Y., & Ming, Z. (2025). A survey of zero-shot object detection. Big Data Mining and Analytics, 8(3), 726-750. https://doi.org/10.26599/BDMA.2024.9020098

[4] Chen, Z., Chen, H., Imani, M., Chen, R., & Imani, F. (2025). Vision language model for interpretable and fine-grained detection of safety compliance in diverse workplaces. Expert Systems with Applications, 265, 125769. https://doi.org/10.1016/j.eswa.2024.125769

[5] Zhang, Z., Yu, Y., Pan, Z., & Antwi-Afari, M. F. (2025). Training-free few-shot construction tool and material detection using pre-trained vision-language model. Computer-Aided Civil and Infrastructure Engineering, 40(30), 6004-6023. https://doi.org/10.1111/mice.70129

[6] Pires, C., Barandas, M., Fernandes, L., Folgado, D., & Gamboa, H. (2020). Towards knowledge uncertainty estimation for open set recognition. Machine Learning and Knowledge Extraction, 2(4), 505-532. https://doi.org/10.3390/make2040028

[7] Zhao, S., Wang, J., Shi, T., & Huang, K. (2024). Contrastive and transfer learning-based visual small component inspection in assembly. Advanced Engineering Informatics, 59, 102308. https://doi.org/10.1016/j.aei.2023.102308

[8] Zhang, J., Qi, Y., & Dai, Y. (2025). Enhancing Object Detection Robustness in Industrial UAVs Through Multimodal Deep Feature Fusion. IEEE Access, 13, 210448-210464. https://doi.org/10.1109/ACCESS.2025.3640280

[9] Gupta, R., Rizve, M. N., Unnikrishnan, J., Tawari, A., Tran, S., Shah, M., ... & Chilimbi, T. (2024, September). Open vocabulary multi-label video classification. In European Conference on Computer Vision (pp. 276-293). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-72933-1_16

[10] Ericsson, L., Gouk, H., Loy, C. C., & Hospedales, T. M. (2022). Self-supervised representation learning: Introduction, advances, and challenges. IEEE Signal Processing Magazine, 39(3), 42-62. https://doi.org/10.1109/MSP.2021.3134634

[11] Zhang, L., Zhang, B., Shi, B., Fan, J., & Chen, T. (2024). Few-shot cross-domain object detection with instance-level prototype-based meta-learning. IEEE Transactions on Circuits and Systems for Video Technology, 34(10), 9078-9089. https://doi.org/10.1109/TCSVT.2024.3395692

[12] Zhang, S., Zhang, Q., Gu, J., Su, L., Li, K., & Pecht, M. (2021). Visual inspection of steel surface defects based on domain adaptation and adaptive convolutional neural network. Mechanical Systems and Signal Processing, 153, 107541. https://doi.org/10.1016/j.ymssp.2020.107541

[13] Tian, Y., Wang, Y., Peng, X., & Zhang, W. (2023). A fault diagnosis method for few-shot industrial processes based on semantic segmentation and hybrid domain transfer learning: Y Tian et al. Applied Intelligence, 53(23), 28268-28290. https://doi.org/10.1007/s10489-023-04979-6

[14] Picard, C., Edwards, K. M., Doris, A. C., Man, B., Giannone, G., Alam, M. F., & Ahmed, F. (2025). From concept to manufacturing: Evaluating vision-language models for engineering design. Artificial Intelligence Review, 58(9), 288. https://doi.org/10.1007/s10462-025-11290-y

[15] Huang, H., Wang, Y., Hu, Q., & Cheng, M. M. (2022). Class-specific semantic reconstruction for open set recognition. IEEE transactions on pattern analysis and machine intelligence, 45(4), 4214-4228. https://doi.org/10.1109/TPAMI.2022.3200384

[16] Zhao, Y., Liu, Q., Su, H., Zhang, J., Ma, H., Zou, W., & Liu, S. (2024). Attention-based multiscale feature fusion for efficient surface defect detection. IEEE Transactions on Instrumentation and Measurement, 73, 1-10. https://doi.org/10.1109/TIM.2024.3372229

[17] Zhang, Y., Kang, B., Hooi, B., Yan, S., & Feng, J. (2023). Deep long-tailed learning: A survey. IEEE transactions on pattern analysis and machine intelligence, 45(9), 10795-10816. https://doi.org/10.1109/TPAMI.2023.3268118

[18] Boult, T. E., Cruz, S., Dhamija, A. R., Gunther, M., Henrydoss, J., & Scheirer, W. J. (2019, July). Learning and the unknown: Surveying steps toward open world recognition. In Proceedings of the AAAI conference on artificial intelligence (Vol. 33, No. 01, pp. 9801-9807). https://doi.org/10.1609/aaai.v33i01.33019801

[19] Amin, S. U., Hussain, A., Kim, B., & Seo, S. (2023). Deep learning based active learning technique for data annotation and improve the overall performance of classification models. Expert Systems with Applications, 228, 120391. https://doi.org/10.1016/j.eswa.2023.120391

[20] Wang, F., Wu, J., Yang, Z., & Song, Y. (2025). Industrial vision inspection using digital twins: bridging CAD models and realistic scenarios. Journal of Intelligent Manufacturing, 36(7), 4963-4975. https://doi.org/10.1007/s10845-024-02485-1

[21] Sardari, S., Fernandes, F., Araya-Martinez, J. M., Zak, J. A., & Roitberg, A. (2025, August). Towards Human-Understandable Visual Recognition for Nonexperts in Industrial Inspection: A Case Study for Car Manufacturing Lines. In 2025 IEEE 21st International Conference on Automation Science and Engineering (CASE) (pp. 279-285). IEEE. https://doi.org/10.1109/CASE58245.2025.11163755

[22] Lu, Z., Sun, H., & Xu, Y. (2023). Adversarial robustness enhancement of UAV-oriented automatic image recognition based on deep ensemble models. Remote Sensing, 15(12), 3007. https://doi.org/10.3390/rs15123007

[23] de Paula Monteiro, R., Lozada, M. C., Mendieta, D. R. C., Loja, R. V. S., & Bastos Filho, C. J. A. (2022). A hybrid prototype selection-based deep learning approach for anomaly detection in industrial machines. Expert Systems with Applications, 204, 117528. https://doi.org/10.1016/j.eswa.2022.117528

[24] Sun, H., He, X., Zhou, J., & Peng, Y. (2023, October). Fine-grained visual prompt learning of vision-language models for image recognition. In Proceedings of the 31st ACM International Conference on Multimedia (pp. 5828-5836). https://doi.org/10.1145/3581783.3612403

[25] Xu, X., Lin, K., Lu, H., Gao, L., & Shen, H. T. (2020, July). Correlated features synthesis and alignment for zero-shot cross-modal retrieval. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 1419-1428). https://doi.org/10.1145/3397271.3401149

[26] Maged, A., & Xie, M. (2022). Uncertainty utilization in fault detection using Bayesian deep learning. Journal of Manufacturing Systems, 64, 316-329. https://doi.org/10.1016/j.jmsy.2022.07.002

[27] Geiß, M., Wagner, R., Baresch, M., Steiner, J., & Zwick, M. (2023). Automatic bounding box annotation with small training datasets for industrial manufacturing. Micromachines, 14(2), 442. https://doi.org/10.3390/mi14020442

[28] Feng, S., Li, B., Yu, H., Liu, Y., & Yang, Q. (2022). Semi-supervised federated heterogeneous transfer learning. Knowledge-Based Systems, 252, 109384. https://doi.org/10.1016/j.knosys.2022.109384

[29] Cohen, J., & Ni, J. (2022). Semi-supervised learning for anomaly classification using partially labeled subsets. Journal of Manufacturing Science and Engineering, 144(6), 061008. https://doi.org/10.1115/1.4052761
[30] Yun, K., Bae, K., & Bae, Y. (2025, August). Task-adaptive open-set detection with prompt-tuned adaptors. In 2025 IEEE International Conference on Advanced Visual and Signal-Based Systems (AVSS) (pp. 1-6). IEEE. https://doi.org/10.1109/AVSS65446.2025.11149799